PREDICTING PROTEIN SECONDARY STRUCTURE USING ARTIFICIAL NEURAL NETWORKS AND INFORMATION THEORY

SAAD OSMAN ABDALLA

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy

Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia

MAY 2005

To my beloved daughters, Raiya and Hiba, and to their late mother

ACKNOWLEDGMENT

I wish to express my deepest and sincere appreciation to my supervisor Professor Dr. Safaai Deris for his guidance, advice, and wise supervision throughout the period of my studentship at the University Technology Malaysia (UTM). His continual support and inspiration, constructive criticisms, and valuable suggestions stormed my brain to navigate the area of Bioinformatics and made this research possible. His high standard of technical ability forced me to challenge a variety of technical problems I faced in this research.

Sincere appreciations and thanks are extended to Professor Zamri bin Mohammed, former Dean of the Faculty of Computer Science and Information Systems (FSKSM), UTM, who taught me several courses in computer sciences, a decade back and then supervised me in my MSC research. Special thanks and appreciation are due to Dr. Abu Baker R. Hussain, Dean of Faculty of Business, Sohar University, Oman, who reviewed the first manuscript of this thesis.

My special thanks and appreciations are due to all my Sudanese friends and collogues at UTM, my brother Abul Nasir, my friend El-Hadi Badawi and families who helped in several ways. Sincere gratitude and appreciations are extended to my colleagues and friends in the Artificial Intelligence and Bioinformatics Laboratory at the Faculty of Computer Science and Information Systems, UTM. Especial thanks here goes to Nazar Zaki and Mohd Saberi.

My ever lasting sincere love and gratefulness go to my wife Nusiabah Ahmed El-Badawi for her patience, inspiration, and continuous support. My love also goes to my lovely daughters Raiya and Hiba and to my mother and my father. Above all, the thanks and gratefulness go to Allah (a.w.j.) for all the good gifts he offered.

ABSTRACT

Large genome sequencing projects generate huge number of protein sequences in their primary structures that is difficult for conventional biological techniques to determine their corresponding 3D structures and then their functions. Protein secondary structure prediction is a prerequisite step in determining the 3D structure of a protein. In this thesis a method for prediction of protein secondary structure has been proposed and implemented together with other known accurate methods in this domain. The method has been discussed and presented in a comparative analysis progression to allow easy comparison and clear conclusions. A benchmark data set is exploited in training and testing the methods under the same hardware, platforms, and environments. The newly developed method utilizes the knowledge of the GORV information theory and the power of the neural network to classify a novel protein sequence in one of its three secondary structures classes. NN-GORV-I is developed and implemented to predict proteins secondary structure using the biological information conserved in neighboring residues and related sequences. The method is further improved by a filtering mechanism for the searched sequences to its advanced version NN-GORV-II. The newly developed method is rigorously tested together with the other methods and observed reaches the above 80% level of accuracy. The accuracy and quality of prediction of the newly developed method is superior to all the six methods developed or examined in this research work or that reported in this domain. The Mathews Correlation Coefficients (MCC) proved that NN-GORV-II secondary structure predicted states are highly related to the observed secondary structure states. The NN-GORV-II method is further tested using five DSSP reduction schemes and found stable and reliable in its prediction ability. An additional blind test of sequences that have not been used in the training and testing procedures is conducted and the experimental results show that the NN-GORV-II prediction is of high accuracy, quality, and stability. The Receiver Operating Characteristic (ROC) curve and the area under curve (AUC) are applied as novel procedures to assess a multi-class classifier with approximately 0.5 probability of one and only one class. The results of ROC and AUC prove that the NN-GOR-V-II successfully discriminates between two classes; coils and not-coils.

ABSTRAK

Projek-projek *genome* yang berskala besar telah menghasilkan jujukan-jujukan protein dalam bentuk struktur pertama yang sangat banyak bilangannya telah menyebabkan teknik-teknik biasa biologi sukar untuk menuntukan struktur 3D dan fungsinya. Peramalan struktur kedua protein diperlukan bagi menentukan struktur 3D protein dan fungsinya. Dalam tesis ini, satu kaedah untuk meramalkan struktur kedua protein telah dicadangkan dan dilaksanakan bersama-sama dengan kaedah-kaedah lain yang berkaitan. Kaedah itu telah dibincangkan dan ditunjukkan di dalam satu analisis perbandingan. Tujuh algoritma dan kaedah bagi peramalan struktur kedua protein telah dibangunkan dan dilaksanakan. Satu set data perbandingan digunakan untuk melatih dan menguji kaedah tersebut. Kaedah yang baru dibangunkan itu adalah menggunakan pengetahuan Teori Maklumat GORV dan Rangkaian Neural untuk mengkelaskan satu jujukan protein baru kepada salah satu daripada 3 kelas stuktur keduanya. NN-GORV-I dibangunkan dan diimplemenkan bagi meramal struktur kedua protein menggunakan maklumat biologi yang disimpan dalam bentuk keladak yang berhampiran dan jujukan-jujukan yang berkaitan. Seterusnya kaedah itu telah diuji dengan kaedah-kaedah lain dan telah mencapai lebih 80% ketepatan. Ketepatan dan kualiti peramalan bagi kaedah itu adalah melebihi 6 kaedah- kaedah lain yang juga telah dibangunkan dan diperiksa dalam penyelidikan ini. Pekali Korelasi Mathews (PKM) telah membuktikan struktur kedua yang telah diramalkan oleh NN-GORV-II adalah sangat berkait rapat dengan keadaan struktur kedua yang telah dicerapkan. Kaedah NN-GORV-II seterusnya diuji dengan menggunakan lima skema potongan DSSP dan disahkan kestabilannya dan boleh dipercayai kebolehannya untuk kerja peramalan tersebut. Satu penambahan ujian bagi jujukan-jujukan yang tidak digunakan dalam prosedur melatih dan menguji dijalankan dan hasil-hasil eksperimennya menunjukkan bahawa peramalan NN-GORV-II adalah berketepatan tinggi, berkualiti dan stabil. Lengkungan *Receiver Operating Characteristic* (ROC) dan *area under curve* (AUC) itu telah diaplikasikan sebagai satu prosedur baru bagi menilai pengkelas pelbagai kelas dengan anggaran kebarangkalian adalah 0.5 bagi satu dan hanya satu kelas. Hasil-hasil bagi ROC dan AUC membuktikan bahawa NN-GOR-V berjaya memisahkan 2 kelas; lingkaran dan bukan lingkaran.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

1D	-	One Dimensional Protein Structure
3D	-	Three Dimensional Protein Structure
HGP	-	Human Genome Project
GenBank	-	Gene Bank
PDB	-	Protein Data Bank
EMBL	-	European Molecular Biology Laboratory
DNA	-	Deoxyribonucleic Acid
RNA	-	Ribonucleic Acid
mRNA	-	Messenger RNA
NMR	-	Nuclear Magnetic Resonance
GOR	-	Garnier-Osguthorpe-Robson
BLAST	-	Basic Local Alignment Search Tool
PSIBLAST	-	Position Specific Iterated Blast
ROC	-	Receiver Operating Characteristic
AUC	-	Area Under Curve
NN-GORV-I	-	Neural Network GOR V Version 1 Prediction Method
NN-GORV-II	-	Neural Network GOR V Version 2 Prediction Method
Q_3	-	Prediction Accuracy of Helices, Strands, And Coils
Q_H	-	Prediction Accuracy of Helix State
Q_E	-	Prediction Accuracy of Strand State
Q_C	-	Prediction Accuracy of Coil State
SOV_3	-	Segment Overlap Measure Of Helices, Strands, And Coils
SOV_H	-	Segment Overlap Measure Of Helix State
SOV_E	-	Segment Overlap Measure Of Strand State
SOV_C	-	Segment Overlap Measure Of Coil State
MCC	-	Matthews Correlations Coefficient
NN	-	Neural Network
CASP	-	Critical Assessment Of Techniques For Protein Structure Prediction

RF	-	Radio Frequency Pulses
CE	-	Combinatorial Extension
FSSP	-	Database F Families Of Structurally Similar Proteins
SCOP	-	Structural Classification Of Proteins
HMMs	-	Hidden Markov Models
FASTA	-	Fast Alignment
GenThreader	-	Genomic Sequences Threading Method
MSA	-	Multidimensional Sequence Alignments
PRINTS	-	Protein Fingerprints
PRODOM	-	Protein Domain
PROF	-	Profile Alignment
PSSM	-	Position Specific Scoring Matrix
PRRP	-	Prolactin-Releasing Peptide
SCANPS	-	Protein Sequence Scanning Package
PHD	-	Profile Network From Heidelberg
DSSP	-	Dictionary Of Protein Secondary Structure Prediction
SAM	-	Sequence Alignment Method
MULTALIGN	-	Multiple Alignment
MULTAL	-	Multiple Alignment
HMMT	-	Hidden Markov Model Training For Biological Sequences
BAliBASE	-	Benchmark Alignments Database
PIM	-	Protein Interaction Maps
ITERALIGN	-	Iteration Alignment
MLP	-	Multi-Layer Perceptron
MI	-	Mutual Information
H	-	α Helix
E	-	β Strand
C	-	Coil
CPU	-	Central Processing Unit
RCSB	-	Research Collaboratory For Structural Bioinformatics
PDB	-	Protein Data Bank
NNSSP	-	Nearest-Neighbor Secondary Structure Prediction
DSC	-	Discrimination Of Protein Secondary Structure Class

3Dee	-	Database Of Domain Definitions (DDD)
CB513	-	Cuff And Barton 513 Proteins
TP	-	True Positive
TN	-	True Negative
FP	-	False Positive
FN	-	False Negative
ANOVA	-	Analysis Of Variance
nr	-	Non Redundant Database
PERL	-	Practical Extraction And Reporting Language
RES	-	Residues
LMS	-	Least Mean Square
SNNS	-	Stuttgart University Neural Network Simulator
ANSI	-	American National Standards Institute
RI	-	Reliability Index
FTP	-	File Transfer Protocol
SPSS	-	Statistical Package For Social Sciences
SAS	-	Statistical Analysis Software
SE	-	Standard Error
PIR	-	Protein Information Resource

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Advances in molecular biology in the last few decades, and the availability of equipment in this field have allowed the increasingly rapid sequencing of considerable genomes of several species. In fact, to date, several bacterial genomes, as well as those of some simple eukaryotic organisms (e.g. yeast) have been completely sequenced. The Human Genome Project (HGP), aimed to sequence all of the human chromosomes, is almost completed with a rough draft announced in the year 2000 (Heilig *et al.*, 2003). Known sequencing databases projects, such as GenBank, PDB, and EMBL, have been growing significantly. This surge and overflow of data and information have imposed the rational storage, organization and indexing of sequence information.

Explaining the tasks undertaken in Bioinformatics field in details might be far beyond this introductory chapter. However, they fall in the creation and maintenance of databases of biological information with nucleic acid or protein sequences cover the majority of such databases. Storage and organization of millions of nucleotides is essential portion in these databases. Designing, developing, and implementing databases access and exchange information between researchers in this field is progressing significantly.

The most fundamental tasks in bioinformatics include the analysis of sequence information which involves the following the prediction of the 3D structure

of a protein using algorithms that have been derived from the knowledge of physics, chemistry and from the analysis of other proteins with similar amino acid sequences. Some researchers refer to this area with the name Computational Biology.

1.2 Protein Structure Prediction

Protein structure prediction is categorized under Bioinformatics which is a broad field that combines many other fields and disciplines like biology, biochemistry, physics, statistics, and mathematics. Proteins are series of amino acids known as polymers linked together into contiguous chains. In a living cell the DNA of an organism encodes its proteins into a sequence of nucleotides (transcribed), namely: adenine, cytosine, guanine and thymine that are copied to the mRNA which are then translated into protein (Branden and Tooze, 1991)

Protein has three main structures: primary structure which is essentially the linear amino acid sequence and usually represented by a one letter notation. Alpha helices, beta sheets, and loops are formed when the sequences of primary structures tend to arrange themselves into regular conformations; these units are known as secondary structure (Pauling and Corey, 1951; Kendrew, 1960). Protein folding is the process that results in a compact structure in which secondary structure elements are packed against each other in a stable configuration. This three-dimensional structure of the protein is known as the protein tertiary structure. However, loops usually serve as connection points between alpha-helices and beta-sheets, they do not have uniform patterns like alpha-helices and beta-sheets and they could be any other part of the protein structure rather than helices or strands (Appendix A).

In the molecular biology laboratory, protein secondary structure is determined experimentally by two lengthy methods: X-ray crystallography method and Nuclear Magnetic Resonance (NMR) spectroscopy method.

Since Anfinsen (1973) concluded that the amino acid sequence is the only source of information to survive the denaturing process, and hence the structured

information must be somehow specified by the primary protein sequence, researchers have been trying to predict secondary structure from protein sequence. Anfinsen's hypothesis suggests that an ideal theoretical model of predicting protein secondary structure from its sequence should exist anyhow.

1.3 Prediction Methods

There are two main different approaches in determining protein structure: a molecular mechanics approach based on the assumption that a correctly folded protein occupies a minimum energy conformation, most likely a conformation near the global minimum of free energy. Potential energy is obtained by summing the terms due to bonded and non-bonded components estimated from these force field parameters and then can be minimized as a function of atomic coordinates in order to reach the nearest local minimum (Weiner and Kollman, 1981; Weiner *et al*., 1984). This approach is very sensitive to the protein conformation of the molecules at the beginning of the simulation.

One way to address this problem is to use molecular dynamics to simulate the way the molecule would move away from that initial state. Newton's laws and Monte Carlo methods were used to reach to a global energy minima. The approach of molecular mechanics is faced by problems of inaccurate force field parameters, unrealistic treatment of solvent, and spectrum of multiple minima (Stephen *et al*., 1990).

The second approach of predicting protein structures from sequence alone is based on the data sets of known protein structures and sequences. This approach attempts to find common features in these data sets which can be generalized to provide structural models of other proteins. Many statistical methods used the different frequencies of amino acid types: helices, strands, and loops in sequences to predict their location. (Chou and Fasman, 1974b; Garnier *et al*., 1978; Lim, 1974b; Blundell *et al*., 1983; Greer, 1981; Warme *et al*., 1974). The main idea is that a

segment or motif of a target protein that has a sequence similar to a segment or motif with known structure is assumed to have the same structure. Unfortunately, for many proteins there is no enough homology to any protein sequence or of known structure to allow application of this technique.

The previous review leads us to the fact that the approach of deriving general rules for predicting protein structure from the existing data sets or databases and then applies them to sequences of unknown structure appears to be promising. Several methods have utilized this approach (Richardson, 1981; Chou and Fasman, 1974a; Krigbaum and Knutton, 1973; Qian and Sejwaski, 1988; Crick, 1989).

Artificial Neural networks have great opportunities in the prediction of proteins secondary structures. These methods are based on the analogy of operation of synaptic connections in neurons of the brain, where input is processed over several levels or phases and then converted to a final output. Since the neural network can be trained to map specific input signals or patterns to a desired output, information from the central amino acid of each input value is modified by a weighting factor, grouped together then sent to a second level (hidden layer) where the signal is clustered into an appropriate class.

Artificial Neural Networks are trained by adjusting the values of the weights that modify the signals using a training set of sequences with known structure. The neural network algorithm adjusts the weight values until the algorithm has been optimized to correctly predict most residues in the training set.

Feedforward neural networks are powerful tools. They have the ability to learn from example, they are extremely robust, or fault tolerant, the process of training is the same regardless of the problem, thus few if any assumptions concerning the shapes of underlying statistical distributions are required. The most promising is that programming artificial neural networks is fairly easy (Haykin, 1999).

Thus, neural networks and specially feedforward networks have a fair chance to well suite the empirical approach to protein structure prediction. In the process of protein folding, which is effectively finding the most stable structure given all the competing interactions within a polymer of amino acids, neural networks explore input information in parallel style.

The GOR method was first proposed by (Garnie *et al.*, 1978) and named after its authors Garnier-Osguthorpe-Robson. The GOR method attempts to include information about a slightly longer segment of the polypeptide chain. Instead of considering propensities for a single residue, position-dependent propensities have been calculated for all residue types. Thus the prediction will therefore be influenced not only by the actual residue at that position, but also to some extent by other neighbouring residues (Garnier and Robson, 1989). The propensity tables to some extent reflect the fact that positively charged residues are more often found in the C-terminal end of helices and that negatively charged residues are found in the N-terminal end.

The GOR method is based on the information theory and naive statistics. The mostly known GOR-IV version uses all possible pair frequencies within a window of 17 amino acid residues with a cross-validation on a database of 267 proteins (Garnier *et al.*, 1996). The GOR-IV program output gives the probability values for each secondary structure at each amino acid position. The GOR method is well suited for programming and has been a standard method for many years.

The recent version GORV gains significant improvement over the previous versions of GOR algorithms by combining the PSIBLAST multiple sequence alignments with the GOR method (Kloczkowski *et al*., 2002). The accuracy of the prediction for the GOR-V method with multiple sequence alignments is nearly as good as neural network predictions. This demonstrates that the GOR information theory based approach is still feasible and one of the most considerable secondary structure prediction methods.

1.4 The Problem

The problem of this research focuses on the protein folding dilemma. The question is how protein folds up to its three dimensional structure (3D) from linear sequences of amino acids? The 3D structure protein is the protein that interacts with each other 3D protein and then produces or reflects functions. By solving the protein folding problem we can syntheses and design fully functioning proteins on a computational machine, a task that may requires several years in the molecular biology labs. A first step towards that is to predict protein secondary structures (helices, strands, and loops). At the time of writing this chapter, the prediction level of protein secondary structures is still at its slightly above the 70% range (Frishman, and Argos, 1997; Rost, 2001; Rost, 2003).

Prediction can not be completely accurate due to the facts that the assignment of secondary structure may vary up to 12% between different crystals of the same protein. In addition, β-strand formation is more dependent on long-range interactions than α-helices, and there should be a general tendency towards a lower prediction accuracy of β-strands than α-helices (Cline *et al.*, 2002).

To solve the above mentioned problems, or in other words to increase the accuracy of protein secondary structure prediction, the hypothesis of this research can be stated as: "construction and designing advanced well organized artificial neural networks architecture combined with the information theory to extract more information from neighbouring amino acids, boosted with well designed filtering methods using the distant information in protein sequences can increase the accuracy of prediction of protein secondary structure".

1.5 Objectives of the Research

The goal of this research is to develop and implement accurate, reliable, and high performing method to predict secondary structure of a protein from its primary

amino acid sequence. However, the specific objectives of this research can be stated in the following points:

a. To analyse and study existing methods developed in the domain of protein secondary structure prediction to help in the development and implementation of a new prediction method.

b. To develop and implement a new accurate, robust, and reliable method to predict protein secondary structure from amino acid sequences.

c. To assess the performance accuracy of the method developed in this research and to compare the performance of the newly developed method with the other methods studied and implemented in this research work.

d. To study the differences between the secondary structure reduction methods and the effects of these methods on the performance of the newly developed prediction method.

e. To carry out blind test on the newly developed method. That is to analyse the output of the newly developed method with respect to an independent data set.

f. To study the performance of the coil prediction of the newly developed method using the ROC curve. This is also to examine the ability of ROC analysis to discriminate between two classes in a multi-class prediction classifier.

1.6 The Scope of This Research

Following the goal and objectives of this study is its scope. Since Bioinformatics is a multi-disciplinary science, the scope of each study must be stated

clearly. The protein sequence data is obtained from the Cuff and Barton (1999) 513 protein database. The data is prepared from the Protein Data Bank (PDB) by Barton's Group and considered as a benchmark sample that represents most PDB proteins. This study focuses on the neural networks and information theory since they are found to be effective for the prediction of protein secondary structure. The output results of the prediction methods are analysed and tested for performance, reliability, and accuracy. The limitation of this research work is the nature of the biological data which needs a great effort of pre-processing before the training and testing stages.

1.7 Organization and Overview of the Thesis

The organization and the flow of the contents of this thesis may be described as follows:

- The thesis begins with Chapter 1 which we are reading now. The chapter explains key concepts, introducing the problem of this research, list the objectives, and determine the scope of this work.

- Chapter 2 reviews and explains the proteins, sequences, and sequence alignments. It also examines amino acids and proteins in terms of their nature, formation, and their importance. The chapter reviews protein homology and homology detection and types of homologies proteins and then explains sequence alignment methods, pair-wise alignment, multiple alignments, as well as profile generation methods.

- The following is Chapter 3 which discusses and overviews protein structure prediction. The generation of profiles that uses the evolutionary information in similar sequences and the multiple sequence alignment methods are thoroughly reviewed in this chapter. This chapter describes the benchmark data sets conventionally used to predict protein structure as well. The chapter also reviews the artificial neural networks and the information theory for prediction of

protein secondary structure with special emphasis to GOR theory. The tools and techniques used in this research as well as prediction performance evaluation procedures are introduced in this chapter..

- Chapter 4 represents a brief and comprehensive methodology of this thesis. The chapter outlines and represents the framework followed in this research to implement the method proposed and developed in this research.

- Chapter 5 represents and explains the modelling of the methodology and algorithms used to develop the new method NN-GORV-I and its advanced version NN-GORV-II. The data set for training and testing the newly developed methods beside the other methods that are implemented in this work was described. The implementation of PSIBLAST program search of the *nr* database to generate multiple sequences which in turns are aligned by the CLUSTALW program is demonstrated in this chapter. The reduction methods used for the secondary structure data and the different statistical analysis and performance tests are demonstrated in this chapter.

- Chapter 6 discusses the results of the seven different prediction methods developed or studied in this research. The Q_3 , the segment overlap (SOV) measure and the Matthews correlations coefficients MCC are discussed and examined in this chapter.

- Chapter 7 discuses the effect of the five eight-to-three secondary structure reduction methods on the newly developed method in this research and trying to judge the argument that the eight-to-three state reduction scheme can alter the prediction accuracy of an algorithm.

- Chapter 8 explores the performance of an independent data set test on the NN-GORV-II method. Few protein targets of CASP3 are

predicted by the newly developed method to judge its performance and quality.

- Chapter 9 introduces the Receiver Operating Characteristics (ROC) analysis and area under curve (AUC) to the newly method which is a multi-class classifier to estimate the prediction accuracy of the coil states.

- Chapter 10 concludes and summarizes this thesis, highlights the contributions and findings of this work, and suggests some recommendations to further extend work.

1.8 Summary

This chapter introduces the problem of predicting protein secondary structure which is the core concern of this thesis. The chapter presents a brief introduction to bioinformatics, proteins, sequences, protein structure prediction. Known methods and algorithms in this domain are briefly introduced and presented. The problem of this research is clearly stated in this chapter and the objectives and scope of this thesis are thoroughly explained. The chapter ends with a description and overview of the organization of the thesis.

REFERENCES

Abagyan, R., Frishman, D. and Argos, P. (1994). Recognition Of Distantly Related Proteins Through Energy Calculations. *Proteins: Structure, Function, and Genetics, Supplement.* 19: 132-140.

Agresti, A. (2002). *Categorical Data Analysis*. 2nd ed. New York, USA: Wiley and Sons.

Alexey, G. M. (1999). Structure Classification-Based Assessment Of CASP3 Predictions For The Fold Recognition Targets. *Proteins: Structure, Function, and Genetics, Supplement*. 3 (1): 88-103.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. (1997). Gapped U BLAST and PSI-BLAST: A New Generation Of Protein Database Search Programs. *Nucleic Acids Research*. 25: 3899-3402.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology*. 215:403-410.

Anderson, T. W. (2003). *An Introduction To Multivariate Statistical Analysis*. 3rd ed. N.Y., USA: Wiley and Sons.

Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C. and Murzin, A. G. (2004). SCOP Database in: Refinements Integrate Structure and Sequence Family Data. *Nucleic Acids Research*. 32:226-229.

Anfinsen, C. B. (1973). Principles That Govern The Folding Of Protein Chains. *Science*. 181: 223-230.

Apostolico, A. and Giancarlo, R. (1998). Sequence Alignment in Molecular Biology. *Journal of Computational Biology*. 5: 173-196.

Attwood, T. K., Beck, M. E., Bleasby, A. J., Degtyarenko, K., Michie A. D. and Parrysmith, D. J. (1997). Novel Developments With The PRINTS Protein Fingerprint Database. *Nucleic Acids Research*. 25: 212-216.

Attwood, T. K., Bradley, P., Flower, D. R., Gaulton, A., Maudling, N., Mitchell, A. L, Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin, A. and Zygouri, C. (2003). PRINTS and Its Automatic Supplement, Preprints. *Nucleic Acids Research*. 31: 400-402.

Aurora, R., Srinivasan, R. and Rose, G. D. (1994). Rules For Alpha-Helix Termination By Glycine. *Science.* 264:1126-1130.

Bairoch, A. and Apweller, R. (1997). The SWISS-PROT Protein Sequence Data Bank and Its Supplement TREMBL. *Nucleic Acids Research*. 25: 31-36.

Bairoch, A. and Boeckmann, B. (1991). The SWISS-PROT Protein-Sequence Data Bank. *Nucleic Acids Research*. 19: 2247-2248.

Bairoch, A., Bucher, P. and Hofmann, K. (1997). The PROSITE Database, Its Status in 1997. *Nucleic Acids Research.* 25: 217-221.

Baldi, P., Brunak, S., Frasconi, P., Pollastri, G. and Soda, G. (1999). Exploiting The Past and The Future in Protein Secondary Structure Prediction. *Bioinformatics*. 15(11): 937–946.

Baldi, P., Brunak, S., Frasconi, P., Pollastri, G. and Soda, G. (2001). *Bidirectional Dynamics For Protein Secondary Structure Prediction, Sequence Learning, Paradigms, Algorithms, and Applications.* 80-104. Springer-Verlag.

Baldi, P., Chauvin, Y., Hunkapillar, T. and Mcclure, M. (1994). Hidden Markov Models Of Biological Primary Sequence Information. *Proceedings of the National Academic of Science*. 91: 1059-1063.

Baldi, P. (1995). Gradient Descent Learning Algorithms Overview: A General Dynamical Systems Perspective. *IEEE Transactions On Neural Networks.* 6(1): 182-195.

Baldi, P. and Brunak, S. (2002). *Bioinformatics: The Machine Learning Approach*. MIT Press.

Baldi, P., Brunak, S., Chauvin, Y., andersen, C. A. F. and Nielsen, H. (2000). Assessing The Accuracy Of Prediction Algorithms For Classification: An Overview. *Bioinformatics*. 16: 412-424.

Barton, G. J. and Sternberg, M. J. E. (1987). A Strategy For The Rapid Multiple Alignment Of Protein Sequences: Confidence Levels From Tertiary Structure Comparisons. *Journal of Molecular Biology*. 198:327-337.

Barton, G. J. (1990). Protein Multiple Sequence Alignment and Flexible Pattern

Matching. *Method Enzymol*. 183: 403-428.

Barton, G. J. (1993). Alscript: A Tool To Format Multiple Sequence Alignments. *Protein Engineering*. 6:37-40.

Bates, P. A. and Sternberg M. J. E. (1999). Model Building By Comparison At CASP3: Using Expert Knowledge and Computer Automation. *Proteins: Structure, Function, and Genetic Supplement*. 3 (1): 47-54.

Benner, S. A. and Gerloff, D. (1991). Patterns Of Divergence in Homologous Proteins As Indicators Of Secondary and Tertiary Structure A Prediction Of The Structure Of The Catalytic Domain Of Protein-Kinases. *Advance in Enzyme Regulation*. 31: 121-181.

Benner, S. A., Badcoe, I., Cohen, M. A. and Gerloff, D. L. (1994). Bona-Fide Prediction Of Aspects Of Protein Conformation Assigning Interior and Surface Residues From Patterns Of Variation and Conservation in Homologous Protein Sequences. *Journal of Molecular Biology*. 235: 926-958.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Ravichandran, V., Schneider, B., Thanki, N., Padilla, D., Weissig, H., Westbrook, J. D. and Zardecki, C. (2002). The Protein Data Bank. *Acta Crystallography*. 58 (6): 899–907.

Bishop, C. (1996). *Neural Networks For Pattern Recognition.* Oxford University Press.

Blundell, T., Sibanda, B. L. and Pearl, L. (1983). Three-Dimensional Structure, Specificity and Catalytic Mechanism Of Renin. *Nature*. 304: 273-275.

Boberg, J., Salakoski, T. and Vihinen, M. (1995). Accurate Prediction Of Protein Secondary Structural Class With Fuzzy Structural Vectors. *Protein Engineering*. 8: 505-512.

Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M. J., Fredholm, H., Lautrup, B. and Petersen, S. B. (1990). A Novel Approach To Prediction Of The 3-Dimensional Structures Of Protein Backbones By Neural Networks. *FEBS. Letters*. 261: 43-46.

Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M., Lautrup, B., Norskov, L., Olsen, O. H. and Petersen, S. B. (1988). Protein Secondary Structure and Homology By

Neural Networks. The Alpha-Helices in Rhodopsin. *FEBS Letters*. 241(1-2): 223-228.

Boscott, P. E., Barton, G. J. and Richards, W. G. (1993). Secondary Structure Prediction For Modelling By Homology. *Protein Engineering*. 6:261-266.

Bowie, J. U., Clarke, N. D., Pabo, C. O. and Sauer, R. T. (1990). Identification Of Protein Folds Matching Hydrophobicity Patterns Of Sequence Sets With Solvent Accessibility Patterns Of Known Structures. *Proteins: Structure, Function, and Genetics, Supplement*. 7: 257-264.

Bowie, J. U., Luthy, R. and Eisenberg, D. (1991). A Method To Identify Protein Sequences That Fold Into A Known 3-Dimensional Structure. *Science*. 253: 164-170.

Bradley, A. P. (1997). The Use Of The Area Under The ROC Curve in The Evaluation Of Machine Learning Algorithms. *Pattern Recognition*. 30 (7): 1145-1159.

Branden, Candtooze, J. (1991). *Introduction To Protein Structure*. Garland Publishing, Inc: New York.

Brenner, S. E. (1996). Molecular Propinquity: Evolutionary and Structural Relationships Of Proteins. University Of Cambridge: PhD Thesis.

Brian, H. (1998). Computing Science: The Invention Of The Genetic Code. *American Scientist.* 86 (1): 9-14.

Briffeuil, P., Baudoux, G., Lambert, C., De Bolle, X., Vinals, C., Feytmans, E. and Depiereux, E. (1998). Comparative Analysis Of Seven Multipl Protein Sequence Alignment Servers: Clues To Enhance Reliability Of Predictions. *Bioinformatics.* 14 (4): 357-66.

Brocchieri, L. and Karlin, S. (1998). A Symmetric-Iterated Multiple Alignment Of Protein Sequences. *Journal of Molecular Biology*. 276(1): 249-64.

Bryant, S. H. and Altschul, S. F. (1995). Statistics Of Sequence-Structure Threading. *Current Opinion in Structural Biology*. 5: 236-244.

Bult, C. J., White, O., Olsen, G. J., Zhou, L. X., Fleischmann, R. D., Sutton, G. G., Blake, J. A., Fitzgerald, L. M., Clayton, R. A., Gocayne, J. D., Kerlavage, A. R., Dougherty, B. A., Tomb, J. F., Adams, M. D., Reich, C. I., Overbeek, R., Kirkness, E. F., Weinstock, K. G., Merrick, J. M., Glodek, A., Scott, J. L., Geoghagen, N. S. M., Weidman, J. F., Fuhrmann, J. L.,

Nguyen, D., Utterback, T. R., Kelley, J. M., Peterson, J. D., Sadow, P. W., Hanna, M. C., Cotton, M. D., Roberts, K. M., Hurst, M. A., Kaine, B. P., Borodovsky, M., Klenk, H. P., Fraser, C. M., Smith, H. O., Woese, C. R. and Venter, J. C. (1996). Complete Genome Sequence Of The Methanogenic Archaeon, Methanococcus Jannaschii. *Science*. 273: 1058-1073.

Burkhard, R. (1999). Twilight Zone Of Protein Sequence Alignments. *Protein Engineering*. 12(2): 85-94.

Burset, M. and Guigo, R. (1996). Evaluation Of Gene Structure Prediction Programs. *Genomics*. 34: 353-367.

Bystroff, C. and Baker D. (1997). Blind Predictions Of Local Protein Structure in Casp2 Targets Using The I-Sites Library. *Proteins: Structure, Function and Genetics Supplement*. 1: 167-171.

Carrington, M. and Boothroyd, J. (1996). Implications Of Conserved Structural Motifs in Disparate Trypanosome Surface Proteins. *Molecular and Biochemical Parasitology*. 81: 119-126.

Chandonia, J. M. and Karplus, M. (1999). New Methods For Accurate Prediction Of Protein Secondary Structure. *Proteins: Structure, Function and Genetics.* 35: 293-306.

Chen, C. P. and Rost, B. (2002). State-Of-The-Art in Membrane Protein Prediction. *Appl. Bioinformatics*. 1: 21-35.

Chothia, C. (1992). Proteins: One Thousand Families For The Molecular Biologist. *Nature*. 357: 543-544.

Chothia, C. and Janin, J. (1975). Principles Of Protein-Protein Recognition. *Nature*. 256: 705-708.

Chothia, C. and Lesk, A. M. (1986). The Relation Between The Divergence Of Sequence and Structure in Proteins. *EMBO Journal*. 5: 823-826.

Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., Sheriff, S., Padlan, E. A., Davies, D. R., Tulip, W. R., Colman, P. M., Alzri, P. M. and Poljak, R. J. (1989). Conformations Of Immunoglobulin Hypervariable Regions. *Nature*. 342: 877-883.

Chou, K. C. and Zhang, C. T. (1994). Predicting Protein-Folding Types By Distance Functions That Make Allowances For Amino-Acid Interactions. *Journal of Biological Chemistry*. 269: 22014-22020.

Chou, K. C. and Zhang, C. T. (1995). Prediction Of Protein Structural Classes. *Critical Reviews in Biochemistry and Molecular Biology*. 30: 275-349.

Chou, P. Y. and Fasman, G. D. (1974b). Prediction Of Protein Conformation. *Biochemistry.* 13: 222-245.

Chou, P. Y. and Fasman, G. D. (1974a). Conformational Parameters For Amino Acids in Helical, Sheet and Random Coil Regions From Proteins. *Biochemistry*. 13: 211.

Chou, P. Y. (1989). Prediction Of Protein Structural Classes From Amino Acid Composition: in: Fasman, G. D. ed. *Prediction Of Protein Structures and The Principles Of Protein Conformation*. Plenum Press. 549-586.

Cline, M. S., Karplus, K., Lathrop, R. H., Smith, T. F., Rogers, R. G., Jr. and Haussler, D. (2002). Information-Theoretic Dissection Of Pairwise Contact Potentials, *Proteins: Structure, Function, and Genetics, Supplement.* 49(1): 7-14.

Crick, F. (1989). The Recent Excitement About Neural Networks. *Nature*. 337: 129-132.

Crooks, G. E. and Brenner, S. E. (2004). Protein Secondary Structure: Entropy, Correlations and Prediction. *Bioinformatics*. 20:1603–1611.

Crooks, G. E., Jason, W. and Steven, E. B. (2004). Measurements Of Protein Sequence Structure Correlations. *Proteins: Structure, Function, and Bioinformatics*. 57:804–810.

Cuff, J. A. and Barton, G. J. (1999). Evaluation and Improvement Of Multiple Sequence Methods For Protein Secondary Structure Prediction. *Proteins: Structure, Function and Genetics*. 34: 508-519.

Cuff, J. A. and Barton G. J. (2000). Application Of Multiple Sequence Alignment Profiles To Improve Protein Secondary Structure Prediction. *Proteins: Structure, Function and Genetics*. 40: 502-511.

Daniel, F., Christian, B., Kevin, B., Arne, E., Adam, G., David, J., Kevin, K., Lawrence, A., Kelley, Robert, M., Krzysztof, P., Burkhard, R., Leszek, R. and Michael, S. (1999). CAFASP-1: Critical Assessment Of Fully Automated Structure Prediction Methods. *Proteins: Structure, Function, and Genetics, Supplement.* 3(1): 209-217.

Defay, T. R. and Cohen, F. E. (1996). Multiple Sequence Information For Threading

Algorithms. *Journal of Molecular Biology*. 262: 314-323.

Depiereux, E., Badoux, G., Briffeuil, P., Reginster, I., De Bolle, X., Vinals, C. and Feytmans, E. (1997). Match-Box-Server: A Multiple Sequence Alignment Tool Placing Emphasis On Reliability. *CABIOS*. 13(3): 249-256.

Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory Of Pattern Recognition*. NY: Springer-Verlag.

Dill, K. A. (1990). Dominanat Forces in Protein Folding. *Biochemistry*. 29: 7133–7155.

Dill, K. A., Bromberg, S., Yue, K. Z., Fiebig, K. M., Yee, D. P., Thomas, P. D. and Chan, H. S. (1995). Principles Of Protein-Folding A Perspective From Simple Exact Models. *Protein Science*. 4: 561-602.

Donnelly, D., Overington, J. P. and Blundell, T. L. (1994). The Prediction and Orientation Of Alpha-Helices From Sequence Alignments The Combined Use Of Environment-Dependent Substitution Tables, Fourier-Transform Methods and Helix Capping Rules. *Protein Eng*ineering. 7: 645-653.

Doolittle, R. F. (1981). Similar Amino-Acid Sequences Chance Or Common Ancestry. *Science*. 214: 149-159.

Dunbrack, R. L. (1999). Comparative Modelling Of CASP3 Targets Using PSI-BLAST Snd SCWRL. *Proteins: Structure, Function, and Genetics, Supplement*. 3(1): 81-7.

Dunbrack, R. L., Gerloff, D. L., Bower, M., Chen, X. W., Lichtarge, O. and Cohen, F. E. (1997). *Meeting Review: The Second Meeting On The Critical Assessment Of Techniques For Protein Structure Prediction (CASP2)*. Asilomar, California.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (2002). *Biological Sequence Analysis: Probabilistic Models Of Proteins and Nucleic Acids*. U.K.: Cambridge University Press.

Eddy, S. R. (1996). Hidden Markov Models. *Current Opinion in Structural Biology*. 6(3): 361-365.

Eddy, S. R. (1998). Profile Hidden Markov Models. *Bioinformatics*. 14(9): 755-63.

Eddy, S.R., Mitchison G. and Durbin R. (1995). Maximum Discrimination Hidden Markov Models Of Sequence Consensus. *Journal of Computational Biology*. 2: 9-23.

Egan, J. P. (1975). Signal Detection Theory and ROC Analysis. *Series in Cognitition and Perceptron*. New York: Academic Press.

Eisenhaber, F., Frommel, C. and Argos, P. (1996). Prediction Of Secondary Structural Content Of Proteins From Their Amino-Acid-Composition Alone: The Paradox With Secondary Structural Class. *Proteins: Structure, Function, and Genetics, Supplement.* 25: 169-179.

Elmasry, N. F. and Fersht, A. R. (1994). Mutational Analysis Of The N-Capping Box Of The Alpha-Helix Of Chymotrypsin Inhibitor-2. *Protein Engineering*. 7: 777-782.

Eyrich, V. A., Przybylski, D., Koh, I. Y. Y., Grana, O., Pazos, F., Valencia, A. and Rost, B. (2003). CAFASP3 in The Spotlight Of EVA. *Proteins: Structure, Function, and Genetics, Supplement*. 53 (6): 548–560.

Farago, A. and Lugosi, G. (1993). Strong Universal Consistency Of Neural Network Classifiers, *IEEE Transactions On Information Theory*. 39: 1146-1151.

Feng, D. F., Johnson, M. S, and Doolittle, R. F. (1985). Aligning Amino Acid Sequences: Comparison Of Commonly Used Methods. *Journal of Molecular*. 21: 112-125.

Feraud, R. and Clerot, R. (2002). A Methodology To Explain Neural Network Classification. *Neural Networks*. 15(2): 237-46.

Ferran, E. A., Pflugfelder, B. and Ferrara, P. (1994). Self-Organized Neural Maps Of Human Protein Sequences. *Protein Science*. 3: 507-521.

Fersht, A. R. (1984). Basis of Biological Specificity. *Trends in Biochemical Science*. 9: 145-147.

Fersht, A. R. (1987). The Hydrogen-Bond in Molecular Recognition. *Trends in Biochemical Science*. 12: 301-304.

Fielding, A. H. and Bell, J. F. (1997). A Review Of Methods For The Assessment Of Prediction Errors in Conservation Presence/Absence Models. *Environmental Conservation*. 24: 38–49.

Fischer, D. and Eisenberg, D. (1996). Protein Fold Recognition Using Sequence-Derived Predictions. *Protein Science*. 5: 947-955.

Fiser, A., Simon, I. and Barton, G. J. (1996). Conservation Of Amino-Acids in Multiple Alignments: Aspartic Acid Has Unexpected Conservation. *FEBS Letters*. 397: 225-229.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., Mckenney, K., Sutton, G., Fitzhugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L. I., Glodek, A., Kelley, J. M., Weidman, J. F., Phillips, C. A., Spriggs, T., Hedblom, E., Cotton, M. D., Utterback, T. R., Hanna, M. C., Nguyen, D. T., Saudek, D. M., Brandon, R. C., Fine, L. D., Fritchman, J. L., Fuhrmann, J. L., Geoghagen, N. S. M., Gnehm, C. L., Mcdonald, L. A., Small, K. V., Fraser, C. M., Smith, H. O. and Venter, J. C. (1995). Whole-Genome Random Sequencing and Assembly Of Haemophilus Influenzae Rd. *Science*. 269: 496-512.

Flockner, H., Braxenthaler, M., Lackner, P., Jaritz, M., Ortner, M. and Sippl, M. J. (1995). Progress in Fold Recognition. *Proteins: Structure, Function, and Genetics, Supplement.* 23: 376-386.

Francesco, V. D., Garnier, J. and Munson, P. J. (1997). Protein Topology Recognition From Secondary Structure Sequences: Application Of The Hidden Markov Models To The Alpha Class Proteins. *Journal of Molecular Biology*. 267(2): 446-463.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J. F., Dougherty, B. A., Bott, K. F., Hu, P. C., Lucier, T. S., Peterson, S. N., Smith, H. O., Hutchison, C. A. and Venter, J. C. (1995). The Minimal Gene Complement Of Mycoplasma Genitalium. *Science*. 270: 397-403.

Freedman, R. B. (1995). The Formation Of Protein Disulfide Bonds. *Current Opinion in Structural Biology*. 5: 85-91.

Frishman, D. and Argos, P. (1995). Knowledge-Based Protein Secondary Structure Assignment. *Proteins: Structure, Function, and Genetics, Supplement*. 23:566-579.

Frishman, D. and Argos, P. (1997). Seventy-Five Percent Accuracy in Protein Secondary Structure Prediction. *Proteins: Structure, Function, and Genetics, Supplement.* 27: 329-335.

Frishman, D. and Argos, P. (1996). Incorporation Of Non-Local Interactions in Protein Secondary Structure Prediction From The Amino-Acid Sequence. *Protein Engineering.* 9: 133-142.

Garnier, J. and Robson, B. (1989). The GOR Method For Predicting Secondary Structures in Proteins. in: Fasman GD, ed. *Prediction Of Protein Structure and The Principles Of Protein Conformation.* New York: Plenum Press. 417-465.

Garnier, J. Gibrat, J. and Robson, B. (1996). GOR Method For Predicting Protein Secondary Structure From Amino Acid Sequence. *Method Enzyme*. 266: 540-553.

Garnier, J., Osguthorpe, D. J. and Robson, B. (1978). Analysis Of The Accuracy and Implications Of Simple Methods For Predicting The Secondary Structure Of Globular Proteins. *Journal of Molecular Biology*. 120: 97-120.

Gibrat, J. F., Garnier, J. and Robson, B. (1987). Further Developments Of Protein Secondary Structure Prediction Using Information-Theory - New Parameters and Consideration Of Residue Pairs. *Journal of Molecular Biology*. 198: 425-443.

Gilbrat, J., Madej, T. and Bryant, S. (1996). Surprising Similarities in Structure Comparison. *Current Opinion in Structural Biology.* 6: 377-85.

Gobel, U., Sander, C., Schneider, R. and Valencia, A. (1994). Correlated Mutations and Residue Contacts in Proteins. *Proteins: Structure, Function, and Genetics, Supplement.* 18: 309-317.

Gotoh, O. (1996). Significant Improvement in Accuracy Of Multiple Protein Sequence Alignments By Iterative Refinement As Assessed By Reference To Structural Alignments. *Journal of Molecular Biology*. 264(4): 823-38.

Gotoh, O. (1999). Multiple Sequence Alignment: Algorithms and Applications. *Advances in Biophysics*. 36(1): 159-206.

Greer, J. (1981). Comparative Model-Building Of The Mamlian Serine Proteases. *Journal of Molecular Biology*. 153: 1027-1042.

Gribskov, M., Luthy, R. and Eisenberg, D. (1990). Profile Analysis. *Method Enzymol*. 183: 146-159.

Gribskov, M., Mclachlan, A. D. and Eisenberg, D. (1987). Profile Analysis Detection Of Distantly Related Proteins. *Proceedings of the National Academic of*

Science. USA. 84:4355-4358.

Grundy, W. N., Bailey, W., Elkan, T. and Baker, C. (1997). Meta-MEME: Motif-Based Hidden Markov Models Of Protein Families. *CABIOS*. 13(4): 397-406.

Gur, D., Rockette, H., and Armfield, D. (2003) Prevalence Effect in A Laboratory Environment. *Radiology*. 228: 10-14.

Han, K. F. and Baker, D. (1995). Recurring Local Sequence Motifs in Proteins. *Journal of Molecular Biology*. 251: 176-187.

Han, K. F. and Baker, D. (1996). Global Properties Of The Mapping Between Local Amino-Acid Sequence and Local-Structure in Proteins. *Proceedings of the National Academic of Science*. USA. 93: 5814-5818.

Hand, D. J. (1997). *Construction and Assignment Of Classification Rules*. NY: John Wiley and Sons.

Hand, D. J. and Till, R. J. (2001). A Simple Generalisation Of The Area Under The ROC Curve For Multiple Class Classification Problems. *Machine Learning*. 45: 171-186.

Hanke, J., Beckmann, G., Bork, P. and Reich, J. G. (1996). Self-Organizing Hierarchical Networks For Pattern-Recognition in Protein-Sequence. *Protein Science*. 5: 72-82.

Hanley, J. A. and Mcneil, B. J. (1983). The Meaning and Use of The Area Under The Receiver Operating Characteristic (ROC) Curve. *Radiology*. 148: 839-43.

Hartl, F. U. (1996). Molecular Chaperones in Cellular Protein-Folding. *Nature***.** 381: 571-580.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. Prentice Hall, Upper Saddle River, NJ.

Heilig, R., Eckenberg, R., Petit J. L., Fonknechten ,N, Da Silva C., Cattolico L., Levy M., Barbe, V., De Berardinis, V., Ureta-Vidal, A., Pelletie,R E., Vico, V., Anthouard, V., Rowen, L., Madan, A., Qin, S., Sun, H., Du, H., Pepin, K., Artiguenave, F, Robert, C, Cruaud, C, Bruls, T., Jaillon, O., Friedlander, L., Samson, G., Brottier, P., Cure, S., Segurens, B., Aniere, F., Samain, S., Crespeau, H., Abbasi, N., Aiach, N., Boscus, D., Dickhoff, R., Dors, M., Dubois, I., Friedman, C., Gouyvenoux, M., James, R., Madan, A., Mairey-Estrada, B., Mangenot, S., Martins, N., Menard, M., Oztas, S., Ratcliffe, A.,

Shaffer, T., Trask, B., Vacherie, B., Bellemere, C., Belser, C., Besnard-Gonnet, M., Bartol-Mavel, D., Boutard, M., Briez-Silla, S., Combette, S., Dufosse-Laurent, V., Ferron, C., Lechaplais, C., Louesse, C., Muselet, D., Magdelenat, G., Pateau, E., Petit, E., Sirvain-Trukniewicz, P., Trybou, A., Vega-Czarny, N., Bataille, E., Bluet, E., Bordelais, I., Dubois, M., Dumont, C., Guerin, T., Haffray, S., Hammadi, R., Muanga, J., Pellouin, V., Robert, D., Wunderle, E., Gauguet, G., Roy, A., Sainte-Marthe, L., Verdier, J., Verdier-Discala, C., Hillier, L., Fulton, L., Mcpherson, J., Matsuda, F., Wilson, R., Scarpelli, C., Gyapay, G., Wincker, P., Saurin, W., Quetier, F., Waterston, R., Hood, L. and Weissenbach, J. (2003). The DNA Sequence and Analysis Of Human Chromosome 14. *Nature*. 421(6923): 601-607.

Henikoff, S and Henikoff, J.G. (1992). Amino Acid Substitution Matrices From Protein Blocks. *Proceedings of the National Academic of Science*. USA 89: 10915-10919.

Henikoff, S., Henikoff, J. G., Alford, W. J. and Pietrokovski, S. (1995). Automated Construction and Graphical Presentation Of Protein Blocks From Unaligned Sequences. *Gene.* 163(2): 17-26.

Henikoff, S. and Henikoff, J. G. (1994). Protein Family Classification Based On Searching A Database Of Blocks. *Genomics*. 19: 97-107.

Henikoff, S. and Henikoff, J. G. (1997). Embedding Strategies For Effective Use Of Information From Multiple Sequence Alignments. *Protein Science.* 6: 698-705.

Henikoff, S. (1996). Scores For Sequence Searches and Alignments. *Current Opinion in Structural Biology*. 6: 353-360.

Higgins, D. G., Thompson, J. D. and Gibson, T. J. (1996). Using CLUSTAL For Multiple Sequence Alignments. *Methods Enzymol.* 266: 383-402.

Hobohm, U., Scharf, M., Schneider, R. and Sander, C. (1992). Selection Of A Represetative Set Of Structures From The Brookhaven Protein Data Bank. *Protein Science* 1: 409-417.

Holley, L. H. and Karplus, M. (1989). Protein Secondary Structure Prediction With A Neural Network. *Proceedings of the National Academic of Science*. USA. 86(1): 152-6.

Holm, L. and Sander, C. (1993). Protein-Structure Comparison By Alignment Of

Distance Matrices. *Journal of Molecular Biology*. 233**:** 123-138.

Hornik, K., Stinchocombe, M., White, H. and Auer, P. (1994). Degree Of Approximation Results For Feedforward Networks Approximating Unknown Mappings and Their Derivatives. *Neural Computation*. 6(6): 1262-1275.

Hornik, K., Stinchocombe, M. and White, H. (1990). Universal Approximation Of Unknown Mapping and Its Derivatives Using Multilayer Feedforward Networks. *Neural Networks*. 3: 535-549.

Huang, X. (1994). On Global Sequence Alignment. *CABIOS*. 10(3): 227-35.

Hubbard, T., Murzin, A., Brenner, S. and Chothia, C. (1997). SCOP: A Structural Classification of Proteins Database. *Nucleic Acids Research*. 25(1): 236-9.

Hubbard, T. J. and Park, J. (1995). Fold Recognition and Abinitio Structure Predictions Using Hidden Markov-Models and Beta-Strand Pair Potentials. *Structure, Function, and Genetics, Supplement.* 23: 398-402.

Hubbard, T. J. P. (1997). New Horizons in Sequence Analysis. *Current Opinion in Structural Biology*. 7: 190-193.

Hudak J. and Mcclure M. A. (1999). A Comparative Analysis Of Computational Motif-Detection Methods. *In Pacific Symposium On Biocomputing*. 138-49.

Hutchinson, E. G. and Thornton, J. M. (1994). A Revised Set Of Potentials For Beta-Turn Formation in Proteins. *Protein Science*. 3: 2207-2216.

Islam, S. A., Luo, J. C. and Sternberg, M. J. E. (1995). Identification and Analysis Of Domains in Proteins. *Protein Engineering*. 8: 513-525.

Jacob, F. (1977). Evolution and Tinkering. *Science*. 196: 1161-1166.

Jimenez, M. A., Munoz, V., Rico, M. and Serrano, L. (1994). Helix Stop and Start Signals in Peptides and Proteins The Capping Box Does Not Necessarily Prevent Helix Elongation. *Journal of Molecular Biology*. 242: 487-496.

Johnson, M. S., Sali, A. and Blundell, T. L. (1990). Phylogenetic-Relationships From 3-Dimensional Protein Structures. *Method Enzymol*. 183: 670-690.

Jones, D. T. (1999b). Genthreader: An Efficient and Reliable Protein Fold Recognition Method For Genomic Sequences. *Journal of Molecular Biology*. 287(4): 797-815.

Jones, D. T. and Thornton, J. M. (1996). Potential-Energy Functions For Threading. *Current Opinion in Structural Biology*. 6: 210-216.

Jones, D. T. (1999a). Protein Secondary Structure Prediction Based On Position-

Specific Scoring Matrices. *Journal of Molecular Biology*. 292: 195-202.

Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992). A New Approach To Protein Fold Recognition. *Nature*. 358: 86-89.

Jones, D.T. and Swindells, M. B. (2002). Getting The Most From PSI-BLAST. *Trends in Biochemistry Science*. 27: 161-164.

Julie, D. T., Frederick, P. and Oliver, P. (1999). Balibase: A Benchmark Alignment Database For The Evaluation of Multiple Alignment Programs. *Bioinformatics*. 15 (1): 87-88.

Julie, D., Thompson, Desmond, G., Higgins, T. and Gibson, J. (1994). CLUSTAL W: Improving The Sensitivity Of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position-Specific Gap Penalties, and Weight Matrix Choice. *Nucleic Acids Research*. 2(22): 4673-4680.

Julie, D., Thompson. F. P. and Oliver, P. (1999). A Comprehensive Comparison Of Multiple Sequence Alignment Programs. *Nucleic Acids Research*. 27(13): 2682-90.

Kabsch, W. and Sander, C. (1983). A Dictionary Of Protein Secondary Structure: Pattern Recognition Of Hydrogen-Bonded and Geometrical Features. *Biopolymers*. 22: 2577-2637.

Kabsch, W. and Sander, C. (1984). On The Use Of Sequence Homologies To Predict Protein-Structure: Identical Pentapeptides Can Have Completely Different Conformations. *Proceedings of the National Academic of Science. USA*. 81: 1075-1078.

Kaur, H. and Raghava, G. (2003). A Neural-Network Based Method For Prediction Of Beta-Turns in Proteins From Multiple Sequence Alignment. *Protein Science*. 12: 923-929.

Kendrew, J. C. Dickerson RE, Strandberg BE, Hart RG, and Davies D.R. (1960). Structure Of Myoglobin. *Nature*. 185: 422-427.

Kevin, K., Christian, B. and Richard, H. (1998). Hidden Markov Models For Detecting Remote Protein Homologies. *Bioinformatics*. 14(10): 846-856.

Kevin, K., Christian, B., Melissa, C., Mark, D., Leslie, G. and Richard, H. (1999). Predicting Protein Structure Using Only Sequence Information. *Proteins: Structure, Function, and Genetics, Supplement*. 3(1): 121-125.

Kevin, K., Kimmen, S., Christian, B., Melissa, C., David, H., Richard, H., Liisa, H.

and Chris, S. (1997). Predicting Protein Structure Using Hidden Markov Models. *Proteins: Structure, Function, and Genetics, Supplement.* 1:134-139.

Kim, H. and Park, H. (2003). Protein Secondary Structure Prediction Based On An Improved Support Vector Machines Approach. *Protein Engineering.* 16(8): 553-60.

King, R. D. and. Sternberg, M. J. E. (1996). Identification and Application Of The Concepts Important For Accurate and Reliable Protein Secondary Structure Prediction. *Protein Science*. 5: 2298-2310.

Klein, P. and Delisi, C. (1986). Prediction of Protein Structural Classes From Amino Acids Sequence. *Biopolymers.* 25: 1659–1672

Kloczkowski, A., Ting, K. L., Jernigan, R. L. and Garnier, J. (2002). Combining The GOR V Algorithm With Evolutionary Information For Protein Secondary Structure Prediction From Amino Aid Sequence. *Proteins: Structure, Function, and Genetics, Supplement*. 49: 154-166

Koretke, K. K., Russell, R. B., Copley, R. R. and Lupas, A. N. (1999). Fold Recognition Using Sequence and Secondary Structure Information. *Proteins: Structure, Function, and Genetics, Supplement*. 3(1): 141-8.

Krigbaum, W. R. and Knutton, S. P. (1973). Prediction of The Amount Of Secondary Structure in A Globular Protein From Its Amino acid Composition. *Proceedings of the National Academic of Science. USA. 70*(10): 2809-2813.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K. and Haussler, D. (1994). Hidden Markov-Models in Computational Biology Applications To Protein Modelling. *Journal of Molecular Biology*. 235: 1501-1531.

Kullback, S., Keegel, J. C. and Kullback, J. H. (1987). *Topics in Statistical Information Theory*. Berlin; New York: Springer-Verlag.

Kulp, D., Haussler, D., Reese, M. G. and Eeckman, F. (1996). A Generalized Hidden Markov Model For The Recognition Of Human Genes in DNA. *Proceedings of the 4th Intelligent Systems for Molecular Biology*.134-142.

Ladunga, I. and Smith, R. F. (1997). Amino Acid Substitutions Preserve Protein Folding By Conserving Steric and Hydrophobicity Properties. *Protein Engineering*. 10: 187-196.

Lathrop, R. H. and Smith, T. F. (1996). Global Optimum Protein Threading With

Gapped Alignment and Empirical Pair Score Functions. *Journal of Molecular Biology*. 255: 641-665.

Lathrop, R. H. (1994). The Protein Threading Problem With Sequence Amino-Acid Interaction Preferences Is NP-Complete. *Protein Engineering*. 7: 1059-1068.

Lattman, E. E. (1995). Protein-Structure Prediction: A Special Issue. *Protein: Structure, Function, and Genetics, Supplement.* 23: 1.

Lawrence, M. C. and Colman.,P.M. (1993). Shape Complementarity at Protein/Protein Interfaces. *Journal. Molecular Biology*. 234: 946–950.

Lesk, A. M., Lo Cont., L. and Hubbard, T. J. P. (2001). Assessment Of Novel Folds Targets in CASP4: Predictions Of Three-Dimensional Structures. Secondary Structures and Inter-Residue Contacts. *Structure, Function, and Genetics, Supplement.* 45(S5): 98–118.

Levitt, M. and Chothia, C. (1976). Structural Patterns in Globular Proteins. *Nature.* 261: 552-557.

Lichtarge, O., Bourne, H. R. and Cohen, F .E. (1996). An Evolutionary Trace Method Defines Binding V Surfaces Common to Protein Families. *Journal of Molecular Biology*. 257: 342-358.

Liisa, H. and Chris, S. (1996). Mapping The Protein Universe. *Science*. 273(5275): 595-603.

Lijmer, J., Mol, B., Heisterkamp, S. (1999). Empirical Evidence Of Design-Related Bias in Studies Of Diagnostic Tests. *Journal of the American Medical Association*. 282: 1061-1066.

Lim, V. I. (1974a). Structural Principles Of The Globular Organisation Of Protein Chains. A Stereochemical Theory Of Globular Protein Secondary Structure. *Journal of Molecular Biology*. 88: 857-872.

Lim, V. I. (1974b). Algorithms For The Prediction Of Alpha-Helical and Beta-Structural Regions in Globular Proteins. *Journal of Molecular Biology*. 88: 873-894.

Lipman, D. J., Altschul, S. F. and Kececioglu, J. D. (1989). A Tool For Multiple Sequence Alignment. *Proceedings of the National Academic of Science.* April USA. 86: 4412-4415.

Lisboa, P. G. J.(Ed) (1992). *Neural Networks: Current Applications*. London: Chapman Hall.

Lise, S. and Jones, D. T. (2005). Sequence Patterns Associated With Disordered Regions in Proteins. *Proteins: Structure, Function, and Bioinformatics*. 58: 144-150.

Maclin, R. and Shavlik, J. W. (1994). Incorporating Advice Into Agents That Learn From Reinforcements. *In Proceedings Of The 12th National Conference On Artificial Intelligence.*

Madej, T., Gibrat, J. F. and Bryant, S. H. (1995). Threading A Database Of Protein Cores. *Structure, Function, and Genetics, Supplement.* 23: 356-369.

Marcella, A., Mcclure, Tatha, K., Vasi and Walter, M. (1994). Comparative Analysis Of Multiple Protein Sequence Alignment Methods. *Molecular Biology and Evolution*. 11(4): 571-592.

Marcella, M., Chris, S. and Pete, E. (1996). Parameterization Studies For The SAM and HMMER Methods Of Hidden Markov Model Generation. *Proceedings of 4th International Conference on Intelligent Systems for Molecular Biology*. 155-164.

Marchler-Bauer, A. and Bryant, S. H. (1997). A Measure Of Success in Fold Recognition. *Trends in Biochemistry Science.* 22: 236-240.

Mark, G. and Michael, L. (1998). Comprehensive Assessment Of Automatic Structural Alignment Against A Manual Standard. The SCOP Classification Of Proteins. *Protein Science*. 7: 445-456.

Marzban, C (2004). A Comment On The ROC Curve and The Area Under It As Performance Measures. *Http://Www.Nhn.Ou.Edu/˜Marzban*.

Matsuo, Y. and Nishikawa, K. (1995). Assessment Of A Protein Fold Recognition Method That Takes Into Account 4 Physicochemical Properties Side-Chain Packing, Solvation, Hydrogen-Bonding, and Local Conformation. *Structure, Function, and Genetics, Supplement.* 23: 370-375.

Matthews, B. B. (1975). Comparison Of The Predicted and Observed Secondary Structure Of T4 Phage Lysozyme. *Biochima et Biophysica Acta.* 405(2): 442-451.

May, A. C. W. and Johnson, M. S. (1994). Protein-Structure Comparisons Using A Combination Of A Genetic Algorithm, Dynamic-Programming and Least-Squares Minimization. *Protein Engineering.* 7: 475-485.

May, A. C. W. and Johnson, M. S. (1995). Improved Genetic Algorithm-Based

Protein-Structure Comparisons Pairwise and Multiple Superpositions. *Protein Engineering.* 8: 873-882.

May, A. C. W. (1996). Pairwise Iterative Superposition Of Distantly Related Proteins and Assessment Of The Significance Of 3-D Structural Similarity. *Protein Engineering*. 9: 1093-1101.

Mcgregor, M. J., Flores, T. P. and Sternberg, M. J. (1989). Prediction Of Beta-Turns in Proteins Using Neural Networks. *Protein Engineering*. 2(7): 521-6.

Metfessel, B. A., Saurugger, P. N., Connelly, D. P. and Rich, S. S. (1993). Cross-Validation Of Protein Structural Class Prediction Using Statistical Clustering and Neural Networks. *Protein Science.* 2: 1171-1182.

Michael, L. (1997). Competitive Assessment Of Protein Fold Recognition and Alignment Accuracy. *Proteins: Structure, Function, and Genetics, Supplement*. 1(1): 92-104.

Michie, A. D., Orengo, C. A. and Thornton, J. M. (1996). Analysis Of Domain Structural Class Using An Automated Class Assignment Protocol. *Journal of Molecular Biology*. 262: 168-185.

Michie, D., Spiegelhalter, D. J. and Taylo, C. C. (1994). *Machine Learning, Neural and Statistical Classification.* Ellis: Horwood.

Mintseris, J. and Weng, Z. (2004). Optimizing Protein Representations With Information Theory. *Genome Informatics.* 15(1): 160-169.

Morgenstern, B., K. Frech, A. Dress and Werner, T. (1998). DIALIGN: Finding Local Similarities By Multiple Sequence Alignment. *Bioinformatics*. 14: 290-294.

Moult, J., Hubbard, T., Bryant, S. H., Fidelis, K., Pedersen, J. T. (1997). Critical Assessment Of Methods Of Protein Structure Prediction (CASP): Round II. *Proteins: Structure, Function, and Genetics, Supplement.* 1(29): 2-6 and113-136.

Moult, J., Hubbard, T., Fidelis, K. and Pedersen, J. (1999). Critical Assessment Of Methods Of Protein Structure Prediction (CASP): Round II. *Proteins: Structure, Function, and Genetics, Supplement.* 3(1): 2-6.

Murzin, A., Brenner S. E., Hubbard, T. and Chothia, C. (1995). SCOP: A Structural Classification Of Proteins Database and The Investigation Of Sequences and Structures. *Journal of Molecular Biology*. 247: 536-540.

Muskal, S. M. and Kim, S. H. (1992). Predicting Protein Secondary Structure-Content A Tandem Neural Network Approach. *Journal of Molecular Biology.* 225: 713-727.

Muskal, S. M., Holbrook, S. R. and Kim, S. H. (1990). Prediction Of The Disulfide-Bonding State Of Cysteine in Proteins. *Protein Engineering. 3*(8): 667-72.

Naderi-Manesh, H., Sadeghi, M., Sharhriar, A., Moosavi and Movahedi, A. A. (2001). Prediction Of Protein Surface Accessibility With Information Theory. *Proteins: Structure, Function, and Genetics, Supplement.* 42: 452–459.

Nagano, K. (1973). Logical Analysis Of The Mechanism Of Protein Folding. *Journal of Molecular Biology*. 75: 401-420.

Nakai, K., Kidera, A. and Kanehisa, M. (1988). Cluster-Analysis Of Amino-Acid Indexes For Prediction Of Protein-Structure and Function. *Protein Engineering*. 2: 93-100.

Nakashima, H., Nishikawa, K. and Ooi, T. (1986). The Folding Type Of A Protein Is Relevant To The Amino-Acid Composition. *Journal of Biochemistry.* 99: 153-162.

Needleman, S. B. and Wunsch, C. D. (1970). A General Method Applicable To The Search For Similarities in The Amino Acid Sequence Of Two Proteins. *Journal of Molecular Biology*. 48: 443-453.

Neuwald, A., Liu, J., Lipman, D. and Lawrenc,E C. E. (1997). Extracting Protein Alignment Models From The Sequence Database. *Nucleic Acids Research*. 25: 1665-1677.

Nielsen, H., Brunak, S. and Von Heijne, G. (1999). Machine Learning Approaches For The Prediction Of Signal Peptides and Other Protein Sorting Signals. *Protein Engineering.*12: 3-9.

Nishikawa, K. and Nogughi, T. (1995). Predicting Protein Secondary Structure Based On Amino Acid Sequence. *Method Enzymol.* 202: 31-44.

Nishikawa, K. and Ooi, T. (1982). Correlation Of The Amino-Acid Composition Of A Protein To Its Structural and Biological Characters. *Journal of Biochemistry*. 91: 1821-1824.

Nishikawa, K., Kubota, Y. and Ooi, T. (1983). Classification Of Proteins Into Groups Based On Amino-Acid Composition and Other Characters (2) Grouping Into Types. *Journal of Biochemistry.* 94: 997-1007.

Norel, R., Lin, S. L., Wolfson, H. J. and Nussinov, R. (1994). Shape Complementarity At Protein-Protein Interfaces. *Biopolymers*. 34: 933-940.

Notredame, C., Holm, L. and Higgins, D. (1998). COFFEE: An Objective Function For Multiple Sequence Alignments. *Bioinformatics.* 14(5): 407-422.

Obuchowski, N. (2000). Sample Size Tables For Receiver Operating Characteristic Studies. *American Journal of Roentgenology*. 175: 603-608.

Olmea, O. and Valencia, A. (1997). Improving Contact Predictions By The Combination Of Correlated Mutations and Other Sources Of Sequence Information. *Folding and Design*. 2: 25-32.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindelis, M. B. and Thornton, J. M. (1997). CATH - A Hierarchic Classification Of Protein Domain Structures. *Structure*. 5(8): 1093-108.

Ouali, M. and King, R. D. (2000). Cascaded Multiple Classifiers For Secondary Structure Prediction. *Protein Science*. 9: 1162-1176.

Pace, C. N., Shirley, B. A., Mcnutt, M. and Gajiwala, K. (1996). Forces Contributing To The Confor- Mational Stability Of Proteins. *Journal of the American Societies for Experimental Biology*. 10(1): 75-83.

Pauling, L. and Corey, R. B. (1951). Configurations Of Polypeptide Chains With Favoured Orientations Around Single Bonds: Two New Pleated Sheets. *Proceedings of the National Academic of Science.* USA. 37**:** 729-740.

Pauling, L. and Corey, R. B. (1951). Configurations Of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets. *Proceedings of the National Academic of Science*. USA. 37**:** 729–740.

Pearson, W. and Lipman, D. (1988). Improved Tools For Biological Sequence Comparison. *Proceedings of the National Academic of Science*. USA 85: 2444-2448.

Pearson, W. R. (1990). Rapid and Sensitive Sequence Comparison With FASTP and FASTA. *Method Enzymol*. 183: 63-98.

Periti, P. F., Quagliarotti, G. and Liquori, A. M. (1967). Recognition Of Alpha Helical Segments in Proteins Of Known Primary Structure. *Journal of Molecular Biology*. 24: 313-322.

Pollastri, G., Przybylski, D., Rost, B., Baldi, P. (2002). Improving The Prediction Of Protein Secondary Structure in Three and Eight Classes Using Recurrent

Neural Networks and Profiles. *Proteins: Structure, Function, and Genetics, Supplement.* 47: 228-235.

Ponder, J. W. and Richards, F. M. (1987). Tertiary Templates For Proteins Use Of Packing Criteria in The Enumeration Of Allowed Sequences For Different Structural Classes. *Journal of Molecular Biology*. 193: 775-791.

Przybylski, D. and Rost, B. (2002). Alignments Grow Secondary Structure Prediction Improves. *Proteins: Structure, Function, and Genetics, Supplement*. 46: 197-205.

Ptitsyn, O. B. (1969). Statistical Analysis Of The Distribution Of Amino Acid Residues Among Helical and Non-Helical Regions in Globular Proteins. *Journal of Molecular Biology*. 42: 501-510.

Qian, N. and Sejnowski, T. J. (1988). Predicting The Secondary Structure Of Globular Proteins Using Neural Network Models. *Journal of Molecular Biology*. 202(4): 865-84.

Rao, S. T. and Rossmann, M. G. (1973). Comparison Of Super-Secondary Structures in Proteins. *Journal of Molecular Biology*. 76: 241-256.

Rice, D. W. and Eisenberg, D. (1997). A 3D-1D Substitution Matrix For Protein Fold Recognition That Includes Predicted Secondary Structure Of The Sequence. *Journal of Molecular Biology*. 267: 1026-1038.

Richard, H. and anders, K. (1996). Hidden Markov Models For Sequence Analysis: Extension and Analysis Of The Basic Method. *Computational Application for* BioScience. 12(2): 95-107.

Richards, F. M. and Kundrot, C. E. (1988). Identification Of Structural Motifs From Protein Coordinate Data: Secondary Structure and First-Level Supersecondary Structure. *Proteins: Structure, Function, and Genetics, Supplement*. 3: 71-84.

Richardson, J. S. (1981). The Anatomy and Taxonomy Of Protein Structure. *Advances in Protein Chemistry*. 34: 168-339.

Richardson, J. S. (1986). The Greek Key Topology As A Favoured Form in Folding and Structure. *Federation Proceedings*. 45: 1829.

Riis, S. K. and Krogh, A. (1996). Improving Prediction Of Protein Secondary Structure Using Structured Neural Networks and Multiple Sequence Alignments. *Journal of Computational Biology*. 3: 163–183.

Rooman, M. J. and Wodak, S. J. (1991). Weak Correlation Between Predictive

Power Of Individual Sequence Patterns and Overall Prediction Accuracy in Proteins. *Proteins: Structure, Function, and Genetics, Supplement.* 9: 69-78.

Rost, B. and Sander, C (1996). Bridging The Protein Sequence-Structure Gap By Structure Predictions. *Annual Review Of Biophysics and Biomolecular Structure*. 25: 113-136.

Rost, B. and Sander, C. (1993). Prediction Of Protein Secondary Structure At Better Than 70% Accuracy. *Journal of Molecular Biology*. 232. 584-599.

Rost, B. (1995). TOPITS: Threading One-Dimensional Predictions Into Three-Dimensional Structures. *Proceedings of the Intelligent System in Molecular Biology*. 314-21.

Rost, B. and Sander, C. (1994). Combining Evolutionary Information and Neural Networks To Predict Protein Secondary Structure. *Proteins: Structure, Function, and Genetics, Supplement.*19: 55-72.

Rost, B. (2001). Review: Protein Secondary Structure Prediction Continues To Rise. *Journal of Structural Biology*. 134: 204–218.

Rost, B. (2003). Neural Networks Predict Protein Structure: Hype Or Hit? Paolo Frasconi ed. *in: Artificial Intelligence and Heuristic Models For Bioinformatics*. CITY:ISO Press. Page

Rost, B. R., Sander, C. and Schneider, R. (1994). Redefining The Goals Of Protein Secondary Structure Prediction. *Journal of Molecular Biology*. 235: 13-26.

Rost, B. and Sander, C. (1993). Prediction Of Protein Secondary Structure At Better Than 70% Accuracy. *Journal of Molecular Biology*. 232: 584–599.

Rost, B., Schneider, R. and Sander, C. (1997). Protein Fold Recognition By Prediction-Based Threading. *Journal of Molecular Biology*. 270: 471-480.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986). Learning Representations By Back-Propagating Errors. *Nature.* 323: 533-536.

Rumelhart, D. E. and Mcclelland, J. L. (1986). *Parallel Distributed Processing: Explorations in The Microstructure Of Cognition.* Cambridge. MA: MIT Press.

Russell, R. B. and Barton, G. J. (1992). Multiple Protein-Sequence Alignment From Tertiary Structure Comparison -- Assignment Of Global and Residue Confidence Levels. *Proteins: Structure, Function, and Genetics, Supplement.* 14: 309-323.

Russell, R. B. and Barton, G. J. (1993). The Limits Of Protein Secondary Structure Prediction Accuracy From Multiple Sequence Alignment. *Journal of Molecular Biology*. 234: 951-957.

Russell, R. B. and Barton, G. J. (1994). Structural Features Can Be Unconserved in Proteins With Similar Folds An Analysis Of Side-Chain To Side-Chain Contacts Secondary Structure and Accessibility. *Journal of Molecular Biology*. 244: 332-350.

Russell, R. B., Copley, R. R. and Barton, G. J. (1996). Protein Fold Recognition By Mapping Predicted Secondary Structures. *Journal of Molecular Biology*. 259: 349-365.

Russell, R. B., Saqi, M. A. S., Sayle, R. A., Bates, P. A. and Sternberg, M. J. E. (1997). Recognition Of Analogous and Homologous Protein Folds: Analysis Of Sequence and Structure Conservation. *Journal of Molecular Biology*. 269: 423-439.

Salamov, A. A. and Solovyev, V. V. (1995). Prediction Of Protein Secondary Structure By Combining Nearest-Neighbour Algorithms and Multiple Sequence Alignments. *Journal of Molecular Biology*. 247: 11-15.

Salamov, A. A. and Solovyev, V. V. (1997). Protein Secondary Structure Prediction Using Local Alignments. *Journal of Molecular Biology*. 268: 31-36.

Sanchez, R. and Sali, A. (1997). Advances in Comparative Protein-Structure Modelling. *Current Opinion in Structural Biology*. 7: 206-214.

Sander, C. and Schneider, R. (1991). Database Of Homology-Derived Protein Structures and The Structural Meaning Of Sequence Alignment. *Proteins: Structure, Function, and Genetics, Supplement*. 9(1): 56-68.

Saqi, M.A., Bates, P. A. and Sternberg, M. J. (1992). Towards An Automatic Method Of Predicting Protein Structure By Homology: An Evaluation Of Suboptimal Sequence Alignments. *Protein Engineering*. 5: 305–311.

Sauder, S. M., Arthur J. W. and Dunbrack, R. L. (2000). Large-Scale Comparison Of Protein Sequence Alignment Algorithms With Structural Alignments. *Proteins: Structure, Function, and Genetics, Supplement.* 40: 6-22.

Schulz, G. E. and Schirmer, R. H. (1979). *Principles Of Proteins Structure*. Springer-Verlag, New York.

Schulz, G. E. (1977). Recognition Of Phylogenetic Relationships From Polypeptide

Chain Fold Similarities. *Journal of Molecular*. 9: 339-342.

Sean, E. (1995). Multiple Alignment Using Hidden Markov Models. in Christopher Railings. *Proceedings in the International Conference of Intelligent Systems for Molecular* Biology. 114-120.

Shannon, C. E. (1948). The Mathematical Theory Of Communications. *Bell System Technical Journal.*

Shindyalov, I. N. and Bourne, P. E. (1998). Protein Structure Alignment By Incremental Combinatorial Extension (CE) Of The Optimal Path. *Protein Engineering*. 11(9): 739-47.

Siddiqui, A. S. and Barton, G. J. (1995). Continuous and Discontinuous Domains An Algorithm For The Automatic-Generation Of Reliable Protein Domain Definitions. *Protein Science*. 4: 872-884.

Siddiqui, A. S., Dengler, U. and Barton, G. J. (2001). 3Dee: A Database Of Protein Structural Domains. *Bioinformatics.* 17: 200-201.

Siegelmann, H. T. (1998). *Neural Networks and Analog Computation: Beyond The Turing Limit*, Boston, Birkhauser.

Siegelmann, H. T. and Sontag, E. D. (1999). During Computability With Neural Networks. *Applied Mathematics Letters.* 4: 77-80.

Sippl, M. J. (1990). Calculation Of Conformational Ensembles From Potentials Of Mean Force An Approach To The Knowledge-Based Prediction Of Local Structures in Globular-Proteins. *Journal of Molecular Biology*. 213: 859-883.

Sippl, M. J., Lackner, P., Domingues, F. S., Prlic, A., Malik, R., andreeva, A. and Wiederstein, M. (2001). Assessment Of The CASP4 Fold Recognition Category. *Proteins: Structure, Function, and Genetics, Supplement.* 5: 55–67.

Sjolander, K., Karplu,S K., Brown, M. P., Hugheym, R., Krogh, A., Mian ,I. S. and Haussler, D. (1996). Dirichlet Mixtures: A Method For Improving Detection Of Weak But Significant Protein Sequence Homology. Computer Application in the Biosciences. 12 (4): 327-345.

Smith, R. F. and. Smith, T. F. (1992). Pattern-Induced Multi-Sequence Alignment (PIMA) Algorithm Employing Secondary Structure-Dependent Gap Penalties For Use in Comparative Protein Modelling. *Protein Engineering*. 5(1): 35-41.

Smith, T. F. (1999). The Art of Matchmaking: Sequence Alignment Methods and

Their Structural Implications. *Structure With Folding and Design.* 7(1): 7-12.

Smith, T. F and Waterman, M. S. (1981). Identification Of Common Molecular Subsequences. *Journal of Molecular Biology*. 147: 195-197.

Sonnhammer., E. L. L. and Kahn., D. (1994). Modular Arrangement Of Proteins As Inferred From Analysis Of Homology. *Protein Science*. 3: 482-492.

Srinivasan, N., Guruprasad, K. and Blundell, T. (1996). Comparative Modelling Of Proteins. in: M. J.Sternberg., ed. *Protein Structure Prediction.* IRL Press. 1-30.

Stephen, R. Holbrook, Steven, M., Muskal and Sung-Hou Kim. (1990). Predicting Protein Structural Features With Artificial Neural Networks. in: Lawrence Hunter ed. *Artificial Intelligence and Molecular Biology*. UK.

Sternberg, M. J. E. and Thornton, J. M. (1976). On The Conformation Of Proteins: The Handedness Of The Beta-Strand - Alpha-Helix - Beta-Strand Unit. *Journal of Molecular Biology*. 105: 367-382.

Swets, J. A., Dawes, R. M and Monahan, J. (2000). Better Decisions Through Science. *Scientific American.* 283: 82-87.

Swets, J. (1988). Measuring The Accuracy Of Diagnostic Systems. *Science*. 240: 1285-1293.

Swindells, M. B. (1995b). A Procedure For The Automatic-Determination Of Hydrophobic Cores in Protein Structures. *Protein Science*. 4: 93-102.

Swingler, K. (1996). *Applying Neural Networks: A Practical Guide.* London: Academic Press.

Tatusov, R., Altschul, S. and Koonin, E. (1994). *Proceedings Of The National Academy Of Sciences Of The United States Of America*. 91(25): 12091-12095.

Taylor, W. R. (1998). Dynamic Sequence Databank Searching With Templates and Multiple Alignments. *Journal of Molecular Biology*. 280(3): 375-406.

Taylor, W. R. and Orengo, C. A. (1989). Protein-Structure Alignment. *Journal of Molecular Biology*. 208: 1-22.

Taylor, W. R. and Thornton, J. M. (1984). Recognition Of Super-Secondary Structure in Proteins. *Journal of Molecular Biology*. 173. 487-514.

Taylor, W. R. (1997). Multiple Sequence Threading: An Analysis Of Alignment

Quality and Stability. *Journal of Molecular Biology*. 269: 902-943.

Thomas, D. J., Casari, G. and Sander, C. (1996). The Prediction Of Protein Contacts From Multiple Sequence Alignments. *Protein Engineering.* 9: 941-948.

Thomas, P. D. and Dill, K. A. (1996). Statistical Potentials Extracted From Protein Structures How Accurate Are They? *Journal of Molecular Biology*. 257: 457-469.

Thompson, J. D., Higgins, D. G. and. Gibson, T. J. (1994). CLUSTAL W: Improving The Sesitivity Of Progressive Multiple Sequence Alignment Through Sequence Weighting, Positions-Specific Gap Penalties and Weigh Matrix Choice. *Nucleic Acids Research*. 22: 4673-4680.

Timothy, L., Bailey and Charles, E. (1994). *Fitting A Mixture Model By Expectation Maximization To Discover Motifs in Biopolymers.* in /SMB-94. 28-36. Menlo Park. CA: AAI/MIT Press.

Tomii, K. and Kanehisa, M. (1996). Analysis Of Amino-Acid Indexes and Mutation Matrices For Sequence Comparison and Structure Prediction Of Proteins. *Protein Engineering.* 9: 27-36.

Valiant, L. (1988). Functionality in Neural Nets, Learning and Knowledge Acquisition. *Proceeding of the American Association for Artificial Intelligent*. 629-634.

Van-Heel, M. (1991). A New Family Of Powerful Multivariate Statistical Sequence-Analysis Techniques. *Journal of Molecular Biology*. 220: 877-887.

Warme, P. K., Momany, F. A., Rumball, S. V., Tuttle, R. W. and Scheraga, H. A. (1974). Computation Of Structures Of Homologous Proteins. Alpha-Lactalbumin From Lysozyme. *Biochemistry. 13*: 768-782.

Weiner, P. K. and Kollman, P. A. (1981). AMBER: Assisted Model Building With Energy Refinement. A General Program For Modeling Molecules and Their Interactions. *Journal of Computational Chemistry.* 2: 287-303.

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. and Weiner, P. (1984). A New Force Field For Molecular Mechanical Simulation Of Nucleic Acids and Proteins. *Journal of American Chemical Societies.* 106: 765-784.

Weiss, S. M. and Kulikowski, C. A. (1991). *Computer Systems That Learn*. Morgan Kaufmann Publishers, Inc, San Mateo. CA.

White, H. (1992). *Artificial Neural Networks: Approximation and Learning Theory*. Blackwell. Oxford.

Woody, R. W. (1995). Circular-Dichroism. *Method Enzymol*. 246. 34-71.

Wu, C. H. and Mclarty, J. W. (2000). Neural Networks and Genome Informatics. *Elsevier Science*.

Wu, C., Whitson, G., Mclarty, J., Ermongkoncha, A. and Chang, T. C. (1992). Protein Classification Artificial Neural System. *Protein Science*. 1: 667-677.

Yang, A. S, Hitz, B. and Honig, B. (1996). Free-Energy Determinants Of Secondary Structure Formation (3) Beta-Turns and Their Role in Protein-Folding. *Journal of Molecular* Biology. 259: 873-882.

Yi, T. M. and Lander, E. S. (1993). Protein Secondary Structure Prediction Using Nearest-Neighbor Methods. *Journal of Molecular* Biology. 232: 1117-1129.

Zachariah, M. A., Crooks, G. E., Holbrook, S. R. and Brenner, S. E. (2005). A Generalized Affine Gap Model Significantly Improves Protein Sequence Alignment Accuracy. *Proteins: Structure, Function, and Bioinformatics*. 58: 329–338.

Zell, A., Mamier, G., Vogt, M., Mache, N., Hubner, R., Doring, S., Herrmann, K. U., Soyez, T., Schmalzl, T., Sommer, T., Hatzigeorgiou, A., Posselt, D., Schreiner, T., Ket., B., Clemente, G. and Wieland. (1998). *The SNNS Users Manual Versi*on 4.1.http://Www.Informatik.Uni-Tuttgart.De/ Ipvr/Bv/Projekte / Snns/Usermanual/User Manual.Html

Zemla, A., Venclovas, C., Fidelis, K. and Rost, B. (1999). A Modified Definition Of SOV: A Segment Based Measure For Protein Secondary Structure Prediction Assessment. *Proteins: Structure, Function, and Genetics, Supplement.* 34: 220-223.

Zhou, G. F., Xu, X. H. and Zhang, C. T. (1992). A Weighting Method For Predicting Protein Structural Class From Amino-Acid-Composition. *European Journal of Neuroscience*. 210: 747-749.

Zou, K. H. (2002). Receiver Operating Characteristic (ROC) Literature Research. *Http://Splweb.Bwh.Harvard.Edu:8000/Pages/ Ppl/Zou/Roc.Html*.

Zvelebil, M. J., Barton, G. J., Taylor, W. R. and Sternberg, M. J. E. (1987). Prediction Of Protein Secondary Structure and Active-Sites Using The Alignment Of Homologous Sequences. *Journal of Molecular* Biology. 195:

957-961.

Zweig, G. and Campbell. C. C. (1993). Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine. *Clinical Chemistry*. 39(4): 561-77.